Data-Parallel Statistical Analysis with R

George Ostrouchov
Oak Ridge National Laboratory

Sean Ahern, Wes Bethel, Jeremy Meredith, Prabhat, Dave Pugmire, Dani Ushizima, and the rest of the VACET team

Data-Parallel Statistical Analysis with R

- Problem
  - Statistical analyses (clustering, regression, extremes modeling, sampling, analysis of variance, etc.) are not available for large data
- Solution
  - Enable R to be used for data-parallel analysis
  - Make data-parallel R available to VisIt visualization
- Impact
  - Enable science to see through the fog of variability in large data sets
  - Facilitate visualization at petascale and beyond
  - Bring half of applied mathematics (statistics) as players to the large data analysis table

Data-Parallel Statistical Analysis with R

- Stakeholders
  - Climate: extremes, uncertainty, and impacts
  - Accelerator: Automatic detection of beam particles in laser-wakefield simulations
  - VisIt: statistical analysis for visualization
- Objectives
  - Enable data-parallel analysis of climate extremes
  - Enable automatic clustering of particles in laser-wakefield simulations
  - Connect R analyses with VisIt visualization

Data Analysts Captivated by R's Power The New York Times **2009**

"... close to 250,000 people work with it regularly."

"Companies as diverse as Google, Pfizer, Merck, Bank of America, the InterContinental Hotels Group and Shell use it."

"R is really important to the point that it's hard to overvalue it," said Daryl Pregibon, a research scientist at Google.

"The great beauty of R is that you can modify it to do all sorts of things," said Hal Varian, chief economist at Google. "And you have a lot of prepackaged stuff that's already available, so you're standing on the shoulders of giants."

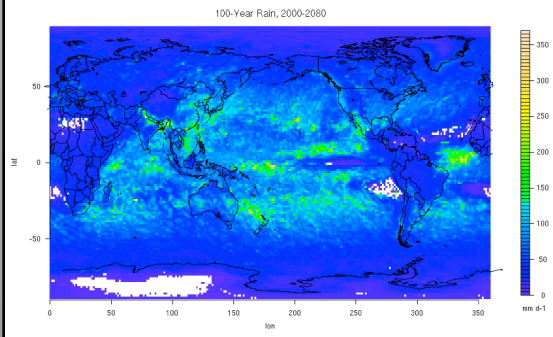
"R has really become the second language for people coming out of grad school now ..." says Max Kuhn, associate director of nonclinical statistics at Pfizer

Milestones

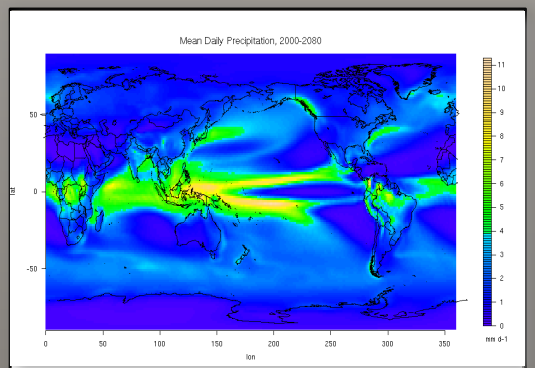
- Parallel R runtime environment: Rmpi+
- Parallel NetCDF data reader
- Data-parallel statistical computing:
  - Map-reduce: histogram, statistical summary, GEV
  - clustering
- Connection to VisIt visualization
  - Parallel NetCDF data writer
- Close the loop with climate and accelerator applications
- Encourage data-parallel programming in the R community

Parallel R Runtime Environment

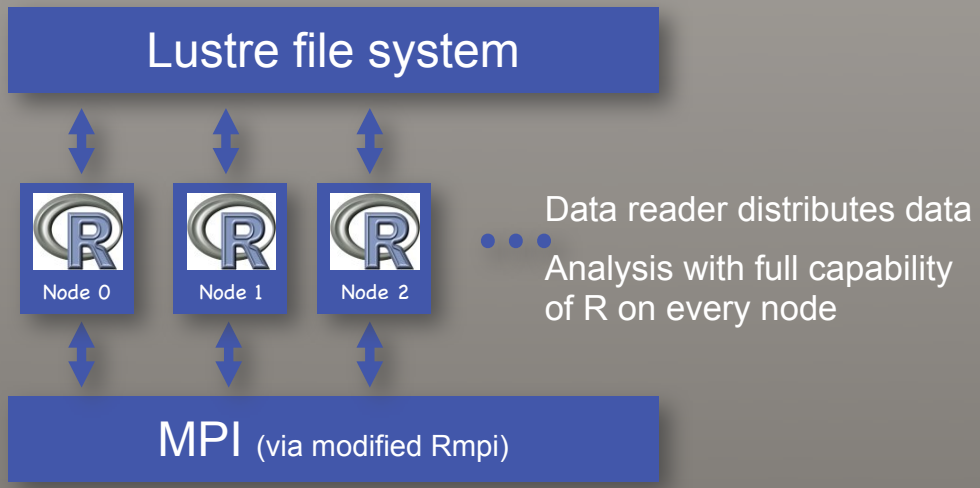
- Batch and interactive
- Basic map-reduce computations: histogram, generalized extreme value (GEV) regression, summary statistics
- k-means clustering with uncertainty - toy version



GEV challenges
- Convergence of 1000s MLE
- Handling failure graciously



Node 0 produces graphics files



Data reader distributes data

Analysis with full capability of R on every node

Accomplishments This Period

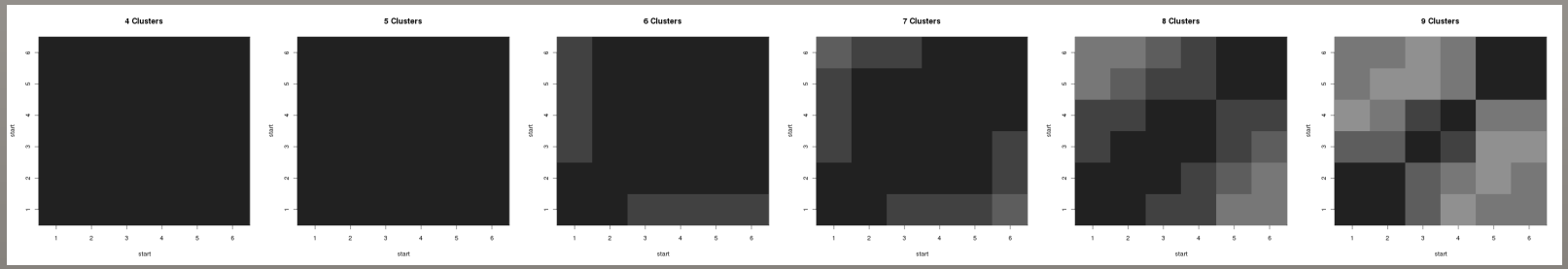
- Model based clustering (k-means) with uncertainty
  - Sampling reduced convergence time to a fraction
  - Good scalability on large vis cluster
- Better understanding of infrastructure that is needed for data-parallel work in R and its connection to VisIt visualization:
  - parallel NetCDF for writing as interface to VisIt
  - interface to BLACS, PBLAS, ScaLAPACK for data-parallel statistical computing
    - e.g. Sliced Inverse Regression: partition/cluster, center, PCA, graphics

k-means Clustering With Uncertainty

- Sampling reduces run time by a factor of 10
- Scalability (3 GB data set) on lens

```
 32 read NetCDF: 18m59.379s  read local: 12m36.411s 11m16.603s 11m23.032s
 64 read NetCDF:  8m16.753s  read local:  5m44.181s  5m46.726s
128 read NetCDF:  5m 8.562s  read local:  3m53.089s
256 read NetCDF:  4m58.799s  read local:  1m50.219s  2m7.691s
```

- Parallel NetCDF writer as interface to VisIt visualization



Cluster coherence for 4 through 9 clusters. Data supports 5 or 6 clusters.

Future Milestones

- Parallel R runtime environment: Rmpi+
- Parallel NetCDF data reader
- Data-parallel statistical computing
  - Map-reduce: histogram, statistical summary
  - clustering
- Connection to VisIt visualization
  - Parallel NetCDF data writer
- Close the loop with climate and accelerator applications
- Data parallel matrix computation
  - BLACS, PBLAS, ScaLAPACK
- Encourage data-parallel programming in the R community
  - JRC 2010, JSM 2010, WIREs: “Data-Parallel Statistical Computing”